

Context-Based Question-Answering Evaluation

Elizabeth D. Liddy, Anne R. Diekema, Ozgur Yilmazel
 Center for Natural Language Processing
 Syracuse University

Question-Answering Environment

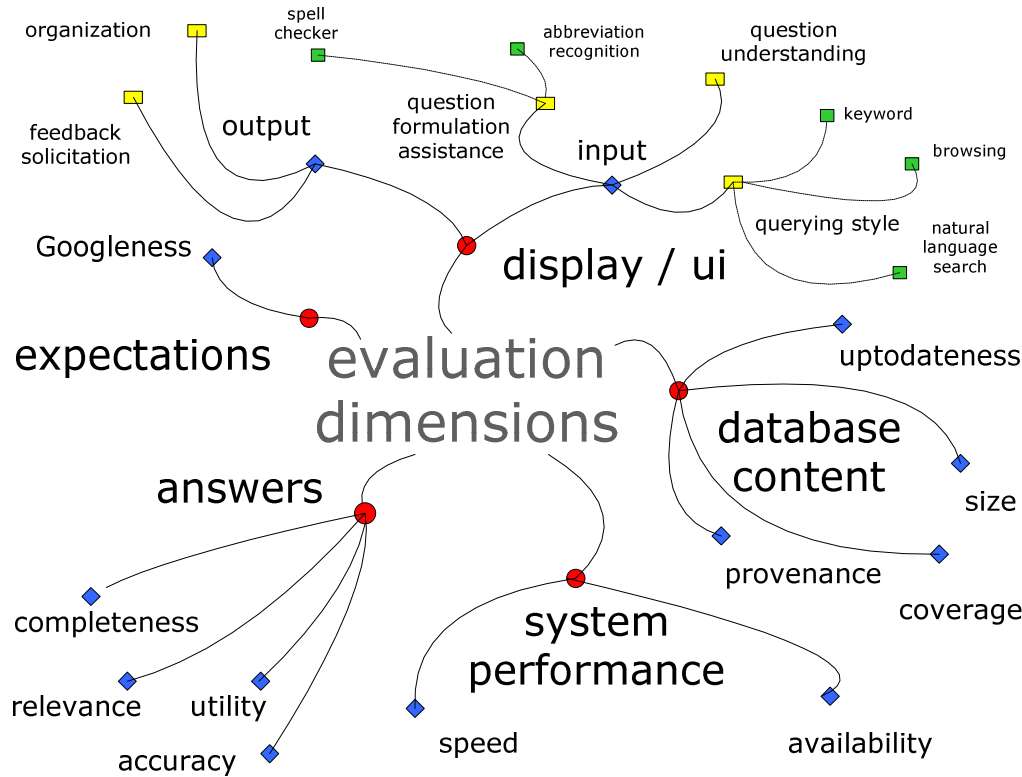
- System: Knowledge Acquisition & Access System (KAAS)
- Domain: Aerospace Engineering – reusable launch vehicles
- Users: undergraduate students
- Collection: textbooks, technical papers and reports, and Website
- Retrieval model: two-stage retrieval model, passage retrieval and domain specific IE
- Answers: pedagogical answer supporting passages
- Funding: NASA, New York State, and AT&T

Situatedness

- QA within a specific:
 - Domain
 - User community
 - Task
- Where QA system must function:
 - In real time, not batch mode
 - With real users' questions
 - Real, not surrogate assessments of relevance
- Therefore, evaluation must reflect the situation

TREC-style Evaluation

- Short, fact based test questions
- Test questions mined from question logs
- Paid assessors
- Adjudicated answers
- Large, public test collection
- System-comparable results
- Actual users not involved



Restricted Domain QA Evaluation

- TREC style evaluation not suitable
- R-D questions are more complex:

Are aerogels rigid enough to sustain the compression inflicted on it by the shell of a sandwich panel-type Thermal Protection System when under the influence of an applied load?

- Provides no sense of the usefulness of the answer

Restricted Domain Evaluation Tasks

- Develop domain specific question set
- Establish correct answers
- Create domain specific test collection
- Create user-based evaluation metrics

Real User-Based Evaluations

- Conducted 2 Surveys of KAAS Users
 - Asking about their experiences with the KAAS
 - 25 to 30 students each for two semesters
 - Open-ended questions
 - Comment on the **usefulness** of KAAS
 - Describe the **quality** of the responses they obtained from the KAAS
- Content analysis of responses by 3 researchers
- Identified 5 major dimensions & 23 minor dimensions in their comments

User-based evaluation metrics

1. System Performance Metrics
 - Answer Return Rate
 - Up-Time
2. Answer Metrics
 - Accuracy or Correctness
 - Completeness
 - Relevance
 - Task Suitability
3. Database Content Metrics
 - Source Quality
 - Coverage
 - Size
 - Recency
4. User Interface Metrics
 - Adaptability
 - Assistance
 - Ease of use
5. Expectation Metrics
 - Expectation satisfaction

Main Evaluation Dimensions

- System Performance
 - "...took so long, so I gave up"
- Answers
 - "...in general what I received was helpful and accurate"
- Database Content
 - "...it needs more documents"
- Display (UI)
 - "...sometimes very good at correcting you to what you need, other times not very good"
- Expectations
 - "...documents in the database were useful, but Google is much faster"